2015

Using Empirical Mode Decomposition to Study Periodicity and Trends in Extreme Precipitation

Noah Pfister
University of Vermont

Using Empirical Mode Decomposition to Study Periodicity and Trends in Extreme Precipitation

A Thesis Presented

by

Noah C. C. Pfister

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fullfillment of the Requirements
for the Degree of Master of Science
Specializing in Mathematics

May, 2015

Defense Date: March 16, 2015
Thesis Examination Committee:

Chris Danforth, Ph.D., Advisor
Lesley-Ann Dupigny-Giroux, Ph.D., Co-Advisor
John Aleong, Ph.D., Chairperson
Jianke Yang, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College

Abstract

Classically, we look at annual maximum precipitation series from the perspective of extreme value statistics, which provides a useful statistical distribution, but does not allow much flexibility in the context of climate change. Such distributions are usually assumed to be static, or else require some assumed information about possible trends within the data. For this study, we treat the maximum rainfall series as sums of underlying signals, upon which we perform a decomposition technique, Empirical Mode Decomposition. This not only allows the study of non-linear trends in the data, but could give us some idea of the periodic forces that have an effect on our series.

To this end, data was taken from stations in the New England area, from different climatological regions, with the hopes of seeing temporal and spacial effects of climate change. Although results vary among the chosen stations the results show some weak signals and in many cases a trend-like residual function is determined.

Acknowledgements

I would like to acknowledge my advisors, Chris Danforth, who introduced me to many of the mathematical concepts discussed in this thesis, as well as Lesley-Ann Dupigny-Giroux, who kindly put up with my antics and was invaluable in finding a focus for my project.

Table of Contents

List of Figures

Chapter 1

Introduction

Due to recent discussions of changing climate, many attempts have been made to look at trends within climatological data. One particularly interesting data set involved in this discussion is the annual (or sometimes daily) maximum taken from precipitation data. Analyzing this data set has important applications in climatology and engineering, since we can use it to determine likely "worst case" scenarios of precipitation (e.g. 100 year floods) over given periods of time. The supposition that the distribution of this data may be changing over time, that the worst rainfall or snowfall of each year may on average be getting worse, presents an important and interesting mathematical quandary. How do we determine if such a trend is present, and how does it affect our estimates of future precipitation?

Typically, annual maximum series (AMS) are studied under the context of extreme value statistics. The data are fit to the general extreme value (GEV) distribution [3], which has probability density functions given by

$$f(y:\mu,\sigma,\kappa) = \frac{1}{\sigma}(1-\kappa\frac{y-\mu}{\sigma})^{(1-\kappa)/\kappa} * \exp(-(1-\kappa\frac{y-\mu}{\sigma})^{1/\kappa}) \tag{1.1}$$

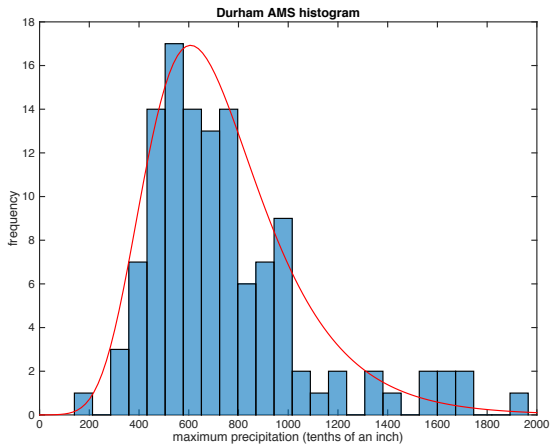


Figure 1.1: A histogram of the AMS data from the ground station in Durham, NH, with a fitted double exponential distribution (extreme value distribution with $\kappa = 0$), acquired from MATLAB's default evd tools [10]. The fitted distribution has parameters $\mu = 607.15$ and $\sigma = 228.20$.

when $\kappa \neq 0$, and

$$f(y : \mu, \sigma) = \exp(-\frac{y-\mu}{\sigma}) * \exp(-\exp(-\frac{y-\mu}{\sigma})) \tag{1.2}$$

when $\kappa = 0$. Here, μ, σ and κ are parameters that govern, respectively, the location, spread, and shape of the distribution. Given the theoretical underpinnings of these distributions, some models will place restrictions on, or even hold fixed, the shape parameter κ.

Recent statistical models have turned to maximum likelihood as a method for fitting distributions to given AMS data. In this method, the likelihood function

$$L(\mu, \sigma, \kappa : y) = f(y : \mu, \sigma, \kappa) \tag{1.3}$$

is determined, which is then maximized over μ, σ, and κ, given the data y, to determine which parameter values best fit the given data. Better yet, if we suspect that the distribution is non-stationary, then we can define each parameter as a function of the time, with its own set of parameters.

The simplest such models assume a linear or quadratic trend for most of their parameters, i.e., setting

$$\mu(t) = \mu_0 + \mu_1 t + \mu_2 t^2 \tag{1.4}$$

and/or

$$\sigma(t) = \sigma_0 + \sigma_1 t + \sigma_2 t^2, \tag{1.5}$$

where setting $\mu_2 = 0$ or $\sigma_2 = 0$ gives us the linear, as opposed to quadratic, model (typically, the κ parameter is left constant, since the shape of extreme value distribution depends more on the shape of the parent distribution, which is not assumed to be changing). Unsurprisingly, increasing the number of variables increases the dimension of the parameter space of the likelihood function, and therefore can lead to less well-defined maxima, meaning that we are far less confident that the parameters we find are indeed the true parameters.

These models are made under the assumption that the true trend can be approximated, at least locally, with simpler functions. One might wonder, however, about the accuracy of these models, and how far in the future we can effectively predict, especially for longer term trends where using a linear approximation may not be as appropriate. Such questions are difficult to answer without some understanding of the true trends (if there are any).

More complex models have been proposed, which consider non-linear trends with varying degrees of flexibility. The idea of using neural networks, essentially considering that the parameters themselves depend on interconnected layers of hidden functions, has been used with a fair bit of success [1, 2]. However, the added complexity of these models somewhat dampens the intuitive, applicable nature of maximum likelihood.

There is another issue at stake: that there may be other time-dependent patterns in the AMS data that may not fit the strictest mathematical definition of trend (that is, as a monotonic function). Due to the inherent cyclic nature of climate phenomena, there exists the possibility that the data will have cyclic properties as well. Such models have been used with fair success for temperature data, for instance, showing periodic, asymmetric trends

during certain seasons, or parts of the year [4]. However, most of these methods already have an assumed periodicity, such as seasonal cycles. In the study of annual maxima, we are left to guess what might be considered the multi-year equivalent of seasons (e.g. teleconnections such as ENSO).

The added complexity of such a model also raises the question of the types of trend. The linear/quadratic trends in classical maximum likelihood engender a changing location or spread to the data, but if we suspect that the data may be at least partially driven by a recurrent pattern, then it is entirely possible that the determining parameters of this recurrent pattern might be changing as well. For instance, it is easy to envision a scenario where the AMS data may be partially driven by a sine-like curve whose amplitude is slowly increasing over time. Physically, such a trend would appear as a changing spread of data, but seeing it instead as an increasingly erratic cycle would not only allow us a better structure with which to make predictions, but give a better glimpse at the effects of climate change, based on how we physically interpret the cyclic pattern in question.

Some more mathematical work must go in to determining if there are appropriate cycles within annual maxima, and seeing if we can form more effective non-linear trend models. Ideally, we want a method that does not assume any particular periods, but rather finds the most appropriate, and perhaps is even capable of producing statistics estimating how closely varying periods fit the trend. Hence, we turn to an emergent method called Empirical Mode Decomposition (EMD). This method does not assume any a-priori knowledge of the dataset, and not only returns any recurring patterns in the data, but has the advantage that the driving mode functions need not be symmetric, or have constant periods (an advantage it has over Fourier analysis or wavelet analysis). Moreover, it allows us to observe possible non-linear trends in the location of the data, and so might give us an idea of how to better model climate change, and to predict how annual precipitation maxima may be changing over time.

Chapter 2

Methods

For this study, a modified version of EMD, Ensemble Empirical Mode Decomposition (EEMD) was used to process each time series. EMD itself is based on the idea that complicated time-series data can be decomposed into a series of imperial mode functions (IMFs), and a left over residual function, by a process first described by *Huang et. al.* [7]. Together, these functions overlay one another to return the original series. Any particular IMF is defined by two properties: first, the number of local extrema and zero crossings can differ by at most one; second, if a continuous "envelope" is defined using the maxima and minima (more on this later), then the mean of the upper and lower envelope must always be zero. Together, these properties give functions that have a strict pattern of alternating local extrema and zero crossings (i.e. between every zero crossing is exactly one extrema, and vice-versa), and that must, in a sense, be centralized around the zero line.

The process by which these IMFs are estimated is fairly intuitive, and moreover, easily modifiable and open to improvement. What follows is the description of the simplest form of EMD, as put forth in Huang's original paper [7]. Each IMF is extracted from the original series using a sifting process which gives the highest frequency IMF present, and subsequent IMFs are found by applying the sifting process to the remaining time series. Consider the case where we have time series $x_0(t)$. First, two "envelope" functions are derived, based on

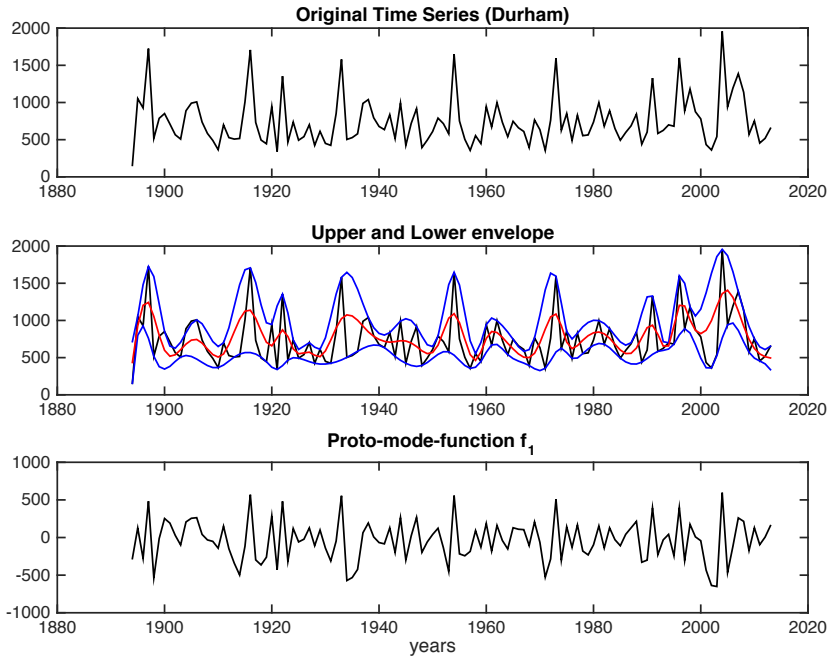


Figure 2.1: Top: The raw time series of annual precipitation maxima from the station in Durham. Middle: The upper and lower "envelope" functions are shown in blue, with the mean function m_1 shown in red. Bottom: The original time series once the mean function has been subtracted off; notice that the residual data now oscillates mainly around 0.

the local extrema of $x_0(t)$. For instance, the upper envelope is the cubic-spline interpolation of the local maxima (the spline need not be cubic [12], but it was chosen for this study as it was easiest to compute in MATLAB, and was the interpolation method used in Huang's proposed method). The lower envelope is derived similarly, using the local minima. The mean of the upper and lower envelopes is then calculated, call it m_1, and then subtracted from the original time series.

Ideally, the remaining time series, $f_1(t) = x_0(t) - m_1(t)$ would be the IMF we are trying to extract, however, the subtraction is enough to create new maxima and minima, and so f_1 may not initially fit the definition of an IMF. Hence, the procedure must be performed on f_1 and each successive function until a stopping criterion is reached. In his initial paper, Huang proposed a Cauchy-type test for convergence, wherein the normalized

squared difference between successive proto-mode-functions is calculated:

$$S_k = \frac{\sum_{t_{start}}^{t_{end}} (f_{k-1}(t) - f_k(t))^2}{\sum_{t_{start}}^{t_{end}} f_{k-1}^2(t)}. \tag{2.1}$$

Once this value drops below some predetermined threshold value, the functions are considered close enough together to be an appropriate approximation of the IMF. Several other similar stopping criteria have been proposed [8]. However, it can be difficult to determine a proper value for this threshold. If it is too large, then f_k may not be a sufficient approximation of the IMF, and if it is too small, then there is the risk of losing possible variations in the amplitude, which may have important meaning in the final interpretation of the IMFs. One remedy was proposed for this, which instead requires that the number of zero crossings and extrema remain the same for some set number of iterations [6]. For this current study, the sifting process was repeated a fixed number of times, which, while very simple, saved quite a bit of time in computation.

Once this highest frequency IMF has been found, it is subtracted from the original time series:

$$x_1(t) = x_0(t) - IMF_1, \tag{2.2}$$

and the process is repeated on x_1 to approximate a second mode function, IMF_2, and then on each subsequent function, making

$$x_k(t) = x_{k-1}(t) - IMF_k. \tag{2.3}$$

This process yields IMFs of decreasing frequency (in the sense of number of zero crossings) until the remaining residue, $r = x_k(t)$, is either a monotonic function, or else does not have enough extrema to properly define envelope functions. Thus, the sum of the IMFs and the residual returns the original time series. Empirically, EMD has been shown to satisfy the requirements of a decomposition method, namely completeness and orthogonality, at

least on the local level [7].

It is easy to see how the IMFs can be useful in interpreting underlying cycles within the physical system of the time series. Moreover, the method is flexible enough to allow functions that are asymmetric, and are not required to have a fixed frequency or amplitude. The residue has multiple interpretations, as it can be seen as an underlying trend, or as the result of potential IMFs with periods larger than the length of the time series (or some combination of both). Similarly to the IMFs, the residual function does not need to be linear, and thus can provide a much more interesting physical meaning.

Several recent improvements have been made to EMD, only some of which were implemented in this project. One problem with EMD on its own is the mixing of mode functions which come close in frequency. This means that small amounts of noise added to the data may lead to very different sets of IMFs. To counter this, an extra layer is added to the EMD, wherein EMD is repeated several times with normally distributed white noise added to the time series. This results in an ensemble of IMF functions, hence this process is known as Ensemble Empirical Mode Decomposition. Corresponding IMFs are averaged to get the ensemble decomposition.

Another useful development in the exploration of EMD is the development of a test for statistical significance of the IMFs. Within any data set there is, presumably, a certain amount of noise, which is almost always assumed to be roughly normally distributed. To study the effects of white noise, Huang analyzed the decomposition of a purely normally distributed time series [15]. What they confirmed was that the EMD method was essentially acting as a dyadic filter bank, returning each IMF with twice the period of the previous one.

Better yet, by analyzing the Fourier spectra of each IMF, they were able to confirm that the mean period of an IMF and the mean energy density are related. Essentially, for

normally distributed data, that

$$\ln \overline{E} + \ln \overline{T} = 0, \tag{2.4}$$

where $\overline{E}$ is the mean energy of any particular IMF, given in the usual Fourier sense as

$$\int |IMF_n(t)|^2 dt, \tag{2.5}$$

and $\overline{T}$ is the mean period of that same IMF, as calculated by counting the number of zero crossings. This conclusion was reached after finding that their Fourier spectra of the IMFs have identical shapes on a logarithmic scale, and therefore

$$\int S_n(\ln T) d \ln T = const., \tag{2.6}$$

where $S_n(\ln T)$ is the Fourier spectrum of the n^{th} IMF of the white noise. By finding an expression for the energy of the n^{th} IMF,

$$\int S_n(\omega) d\omega \tag{2.7}$$

(where $S_n(\omega)$ is the Fourier spectrum of the IMF as a function of the frequency ω), and then applying some clever changes of variables, they were able to show that the product of the energy and the average period is equal to (2.6), and since, for normalized white noise, we can assume that the constant is 1, then we get the relation expressed in (2.4) [15].

If, by the central limit theorem, we can assume that the white noise of a time series is normally distributed around the base time series, then we get that the mean energy will have a χ^2 distribution with degrees of freedom as determined by the expected energy of each IMF, predicted by using equation (2.4). Hence, for an IMF with any given mean period, we can determine its statistical significance by looking at its energy density and seeing into what percentile it falls on the relevant χ^2 distribution [8, 15].

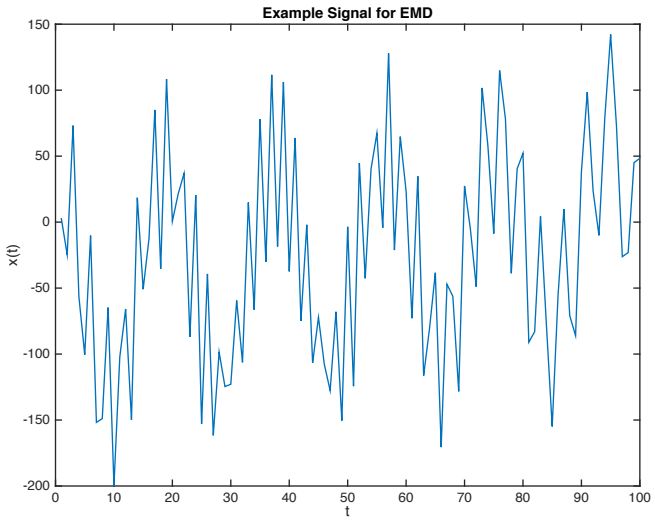


Figure 2.2: A generated signal to test the effectiveness of EMD.

As a demonstration of how EMD is used, and to see what results look like before the methods are applied to AMS data, suppose we were to generate a time series from predetermined mode functions, and see how well we can recover said signals. For this purpose, we construct a time series generated by

$$x(t) = 75\cos(t^2/66 + 2t) + (90 - t/5)\cos(t/5) + 20\cos(t/98 + 10) \tag{2.8}$$

where t is each whole number between 1 and 100 (standing in for the discrete years we will be dealing with in AMS data). This generating series was chosen to have several different frequency modes, with subtly changing frequencies and amplitudes, in the hopes of seeing how EMD recovers these effects as well.

We apply simple EMD to this series, and compute the mean period and energy density to test statistical significance. There are several interesting things to note about the conclusions. First, the decomposition was able to find the middle period mode function $((90 - t/5)\cos(t/5))$, although it appears that part of the signal was transferred to the

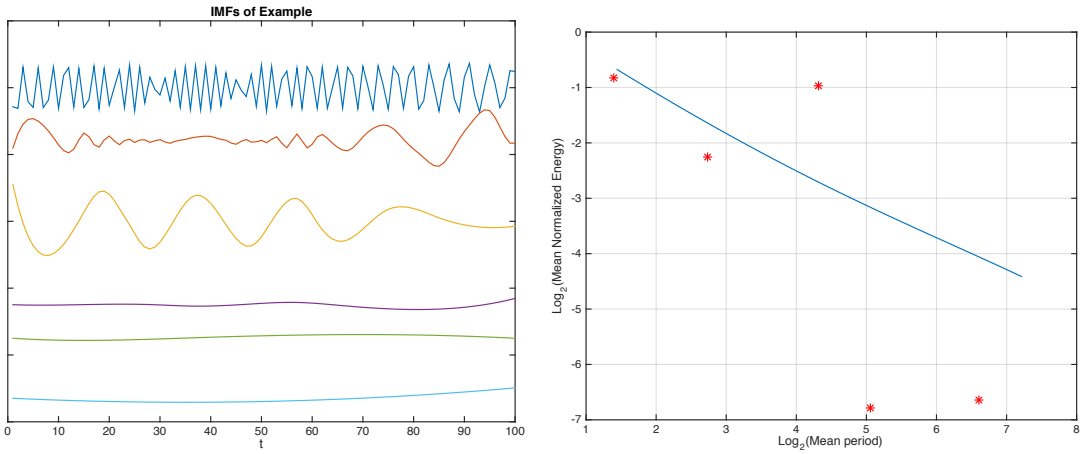


Figure 2.3: The IMFs generated from our example problem, along with the relative significance of each (with marks representing each of the IMFs). Notice that the recovered signals do not exactly match the generating signals, though they are remarkably close. The blue line in the significance plots represents 95% significance, where each red mark represents one of the IMFs. If an IMF is significant, then corresponding mark will be above the blue curve.

second IMF. Such problems are not uncommon, and are likely due to the mode mixing of various IMFs.

Another aspect to note is that the the highest frequency mode ($75\cos(t^2/66 + 2t)$) was also captured, and even displays the changing frequency of the original function, despite that it was not determined to be significant. Since, for the normally distributed data on which the significance test is based, the mean energy is very high for lower period IMFs, then it is often much harder to determine significance for low periods.

One other note of interest, and in fact one of the most remarkable outputs of EMD analysis, is the presence of the largest period mode function ($20\cos(t/98 + 10)$), intended to represent an underlying trend-like function. The function is not captured perfectly, since the trend was made intentionally subtle, and notably breaks down a little toward the endpoints, but it did manage to find a generally upward residual function.

Chapter 3

Data and Results

For this study, data were taken from 10 different stations spread across Vermont and New Hampshire (see appendix for original time series). This particular study region was chosen in part because of the locality, but also in the hopes that the varied topography and climate would yield particularly interesting results. Ground station data were acquired from the National Climate Data Center's online data site [11], and chosen to fit several criteria. First, the data were required to span about 100 years (preferably more). Although we expected that any mode functions we might detect would probably have a period much shorter than this timescale, we must also have a series long enough to establish a pattern.

Second, the data had to be relatively complete. While a single missing data point is not such a problem, since we could extrapolate some approximation from the surrounding points, larger gaps present a problem. Since EMD relies almost entirely on the spline of the local maxima and minima, then we have cause to be wary of the endpoints in our results. The problems are compounded if we consider that one of these spline functions would extend over these large gaps in the data and so might mis-represent the shape of the envelope at nearby locations. Dividing the data at the gap and processing each piece separately is one possible solution, but often reduces the time series into two sections that are no longer long enough to get useful IMFs.

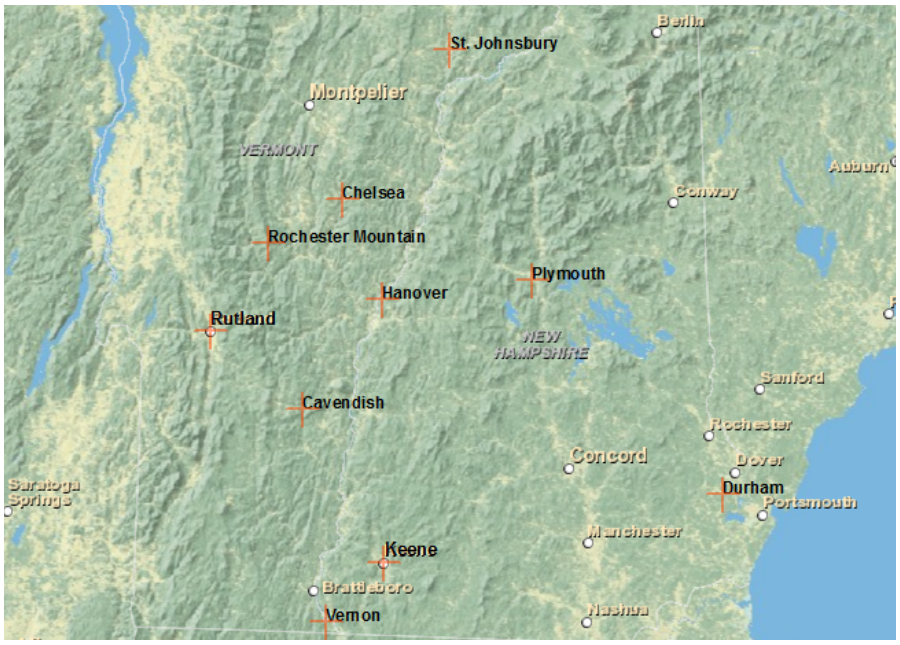


Figure 3.1: A map of northern New England showing ground stations used in this study (marked in red). Base layer data were extracted from those provided from ArcGIS Online

To perform our EEMD operations, we opted to use code provided by *Huang and Wu* [5], implemented in MATLAB [10] (though several iterations were performed by hand, including those for the example problem stated earlier, so as to confirm the validity of the code). EEMD was performed with 100 iterations, with a noise with standard deviation 0.2 of the original series (as per Huang's initial examples). Results were then plotted and run through our test for significance (a few examples are shown here, see Appendix for remaining IMFs).

For the most part, very few significant IMFs were found. Notable exceptions were Durham, NH, whose first and second mode functions were at least 95% significant with mean periods of 2.629 years and 6.000 years, respectively. Keen, NH, similarly had a significant first IMF, with mean period 2.811 years (although the large spike toward the end of the Keene time series made several of its IMFs unusually shaped, possibly as a result of the spline interpolating functions crossing). Significant mode functions were also found

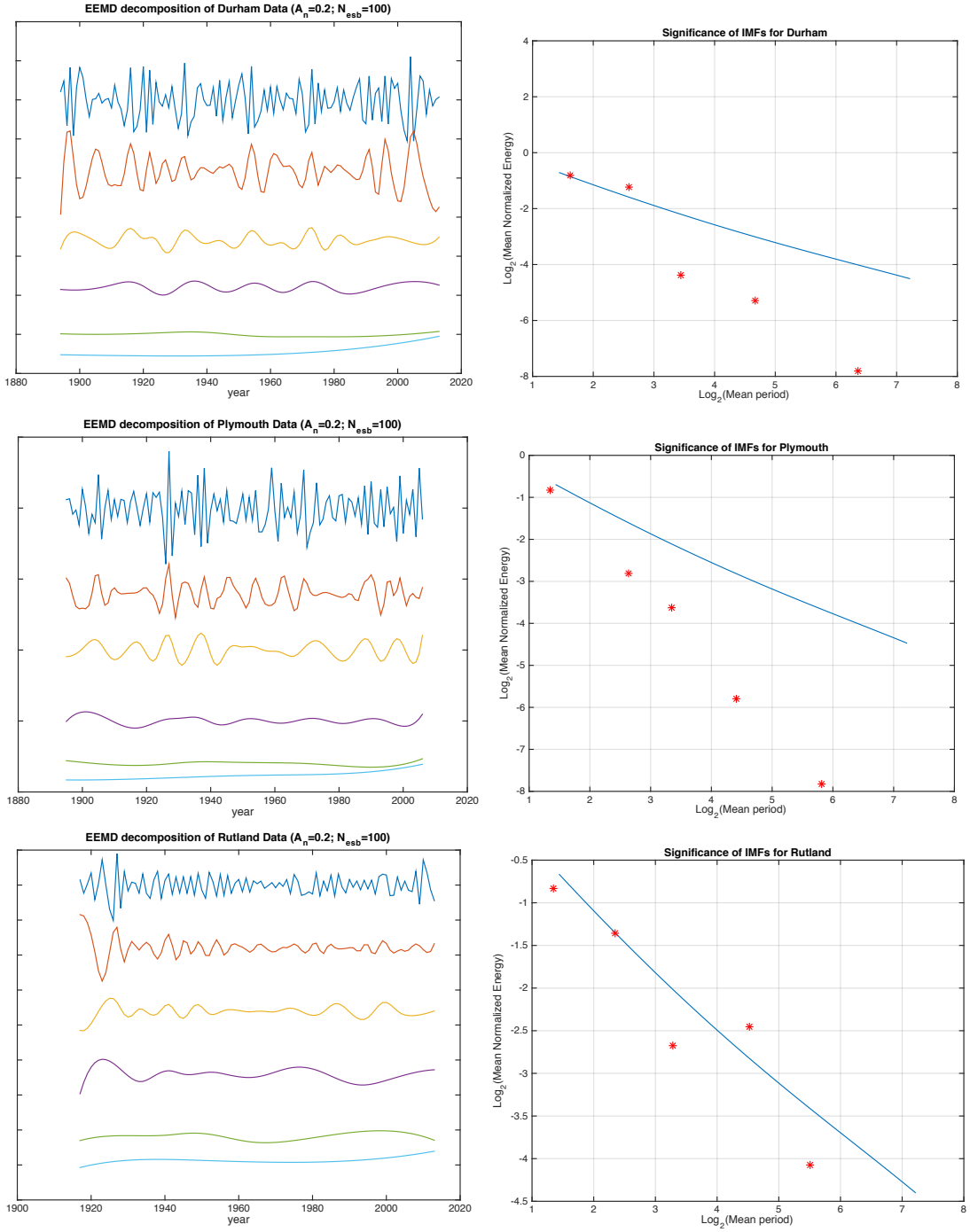


Figure 3.2: The IMFs (and residues on on the last row of each plot) generated by data from stations at Durham and Plymouth, NH, (in descending order of frequency), along with the plots of their respective significance.

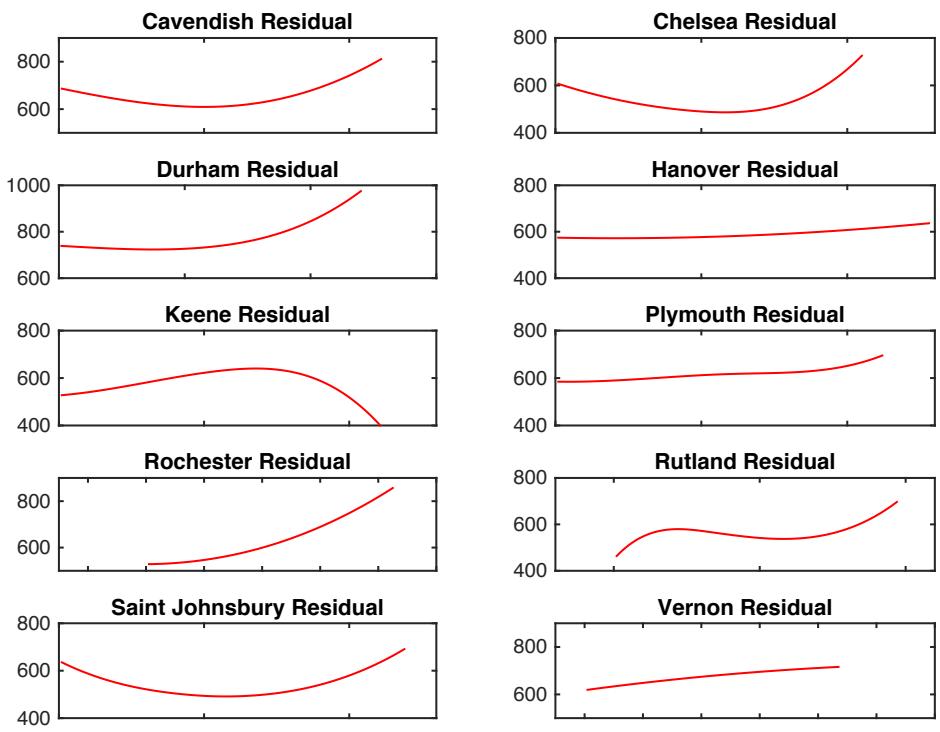


Figure 3.3: The residual functions from each of the ten stations, in units of tenths of an inch.

at Rutland, VT, with mean periods 5.532 years and 20.494 years. However, it should be noted that the standard of 95% was set fairly low, and that none of the IMFs exceeded the somewhat stricter standard of 99% significance.

A slightly more interesting result is the presence of trend-like residual functions. Unfortunately, there is no suitable significance test for the residuals, since they are not typically the focus of EMD study. While slight, they are noticeable compared to the scale of the other IMFs.

Chapter 4

Conclusions

Of the 10 time series tested, none showed anything much more patterned than randomly distributed data around a trend-like function. There could be several reasons for this. The AMS data for precipitation are notoriously noisy, either due to error in measurement, or to the fairly noisy nature of rainfall data in general. While it should not have a significant effect, especially for a data set as abstracted as annual extrema, it is possible that due to changes in station location or in the method by which the data are acquired (e.g. improvements in rain-guage technology), the quality of data may deteriorate as we look further back in time. This, however, is for the moment mere speculation.

Another, more likely, explanation is that any true IMFs are too subtle to be detected by this method. Although we may intuitively suppose that there are cyclic forces underlying the data, it is clear from our results that their effect is just too small compared to very large random noise. What this means, essentially, is that from any given year, we cannot easily predict whether there will be future periods of better or worse precipitation maxima, or at least for periods of less than the length of our time series.

Any information pertaining to longer period patterns would, however, be found in the residual functions, along with information on any trends. Although EMD is certainly not the most efficient route to finding a non-linear trend, our results give us a glimpse at why

the study of trend is difficult. Studies looking at long term change in precipitation often have trouble finding trends in AMS data [13], and our results may demonstrate why. Take, for example, the time series for Cavendish, VT. Trend analysis would likely find a weak trend, if any, when applied to the whole series, where a much stronger trend would be present if, say, only the most recent half of the time series was considered.

Conversely, we could see the curve of the residual as a brief section of a larger IMF-like function, that our series is simply not long enough to fully capture. In this case, what a simpler test may interpret as a trend, may be entirely due to what section of the curve is represented in the time series. Depending on the interpretation of the residual functions, it is possible that a longer cycle, or even some non-cyclic outside force, may confound our search for a trend. Without longer time series, or some other form of outside information, it is impossible to reach a conclusion.

The important point that can be made from this analysis is that having more time series data requires more complicated models to make accurate predictions, at least in the context of climate change. Consider that one of the most commonly used statistics in the study of extreme values, particularly for precipitation, is the return period, that is, given a particular maximum precipitation, over what period of time will that maximum or greater recur (or, directly related, given an interval of time, what is the worst daily precipitation we would see in that time). Such information can be easily be acquired from a stationary distribution, or even from a time dependent distribution [3, 9]. There is an important question, however, in how we would incorporate the non-linear residual functions.

Coming up with a general model directly from the residual functions could prove difficult, especially since they seem to vary by station. What the IMF functions do allow us to do is to study the predictive power of simpler models over different time scales. For instance, a model that has the location parameter μ changing linearly over time might be fairly accurate in sections of the series where the residual functions are relatively straight, such as the earlier

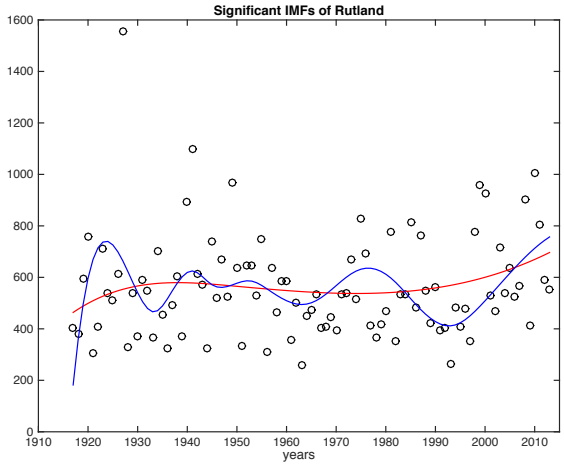


Figure 4.1: The data from Rutland, VT, plotted against the found residual (in red), and the most significant IMF with added trend (in blue). Notice that even though this IMF was significant, it does not do a great job of predicting the important parts of the data, that is, the upper peaks.

parts of the time series for Durham NH, or Keene, NH. On the other hand, a linear model would not predict so well during periods where the residual is highly curved. Perhaps, given the results of EMD and some statistical work, it would be possible to determine a way of calculating return periods of precipitation which uses a linear model, but only takes the amount of data necessary, so as to preserve the linear approximation at the local level. Finding a balance with how much information we need to make an accurate prediction would be key to such research.

There are several other further improvements that could be made to our methods. One of the assumptions made in our EMD analysis, particularly in determining the significance of each IMF, was that the data set is roughly normally distributed. However, as previously stated, AMS data are commonly accepted as having a GEV distribution, and one further area of study would be in determining how strongly this difference effects the results of EMD.

One key feature of the GEV distribution is its skewed shape leading to scattered higher outcome values. One of the biggest concerns in the application of AMS data analysis (or

indeed any study of extreme values) is the the occurrence of such points, the extremes of the extremes you might say, and this is an aspect in which we would hope our model would be particularly accurate, if possible. If, say such spikes rise and fall in a periodic fashion, then we would hope that EMD analysis would reflect this information. However, even for the time series where we found significant IMFs, we can see that the EMD model does a very poor job capturing the peaks. This is one area in which improvements could be made.

Another area of future research lies in dealing with missing data in the context of EMD. This became a problem in our study, where stations frequently shut down part of their data collection, or changed location, leading to large gaps in the data, sometimes of more than ten or twenty years. Barring discussions of the reliability of using such data as a full series, we could ask the question of how we would recover useful IMFs. Conventional wisdom suggests that we could simply find a neutral set of values, derived from the surrounding data, such that impact to the overall pattern is minimized. Problems arise, however, when we proceed through the sifting portion of EMD. Do we allow the points to change, and therefore possibly become local maxima or minima, or do we keep them neutral and let the spline function be interpolated over these large gaps? Unfortunately, there was not enough time to fully explore these problems.

Tied to the missing data problem is the problem of endpoints. In his first paper, Huang noted that the spline interpolation tends to break down at both ends of the time series [7]. Classical solutions to this involve setting end conditions for the spline functions so that they remain relatively well behaved. *Huang et. al.* has also proposed a method by which a "frame" of extra points was created around the endpoints of the given time series, in such a way that as much information as possible is extracted from the true data. Without a better theoretical understanding off the EMD process, which is still very much under exploration, it is very difficult to establish whether such methods are truly better at pulling information from every single part of that data set.

In a recent paper on the Hilbert-Huang transform, of which EMD is a part, *Huang et. al.* noted several open problems [8]. Criteria for decomposition methods often require that the output be unique. That is, in the context of EMD, that the IMFs we find for any given time series are the only mode functions that can be used to recover the original function, and ideally, with the fewest number of IMFs needed. It remains to be shown that EMD is able to give unique solutions. Such questions are very difficult to answer without a more rigorous definition of an intrinsic mode function (perhaps something along the lines of defining them as $a(t)\cos(\theta(t))$ where $\theta(t)$ is a piecewise smooth increasing function, and $a(t)$ is smooth [14]). Moreover, if they are not unique, can we optimize our EMD procedure so that the IMFs are a unique "best fit" for the series?

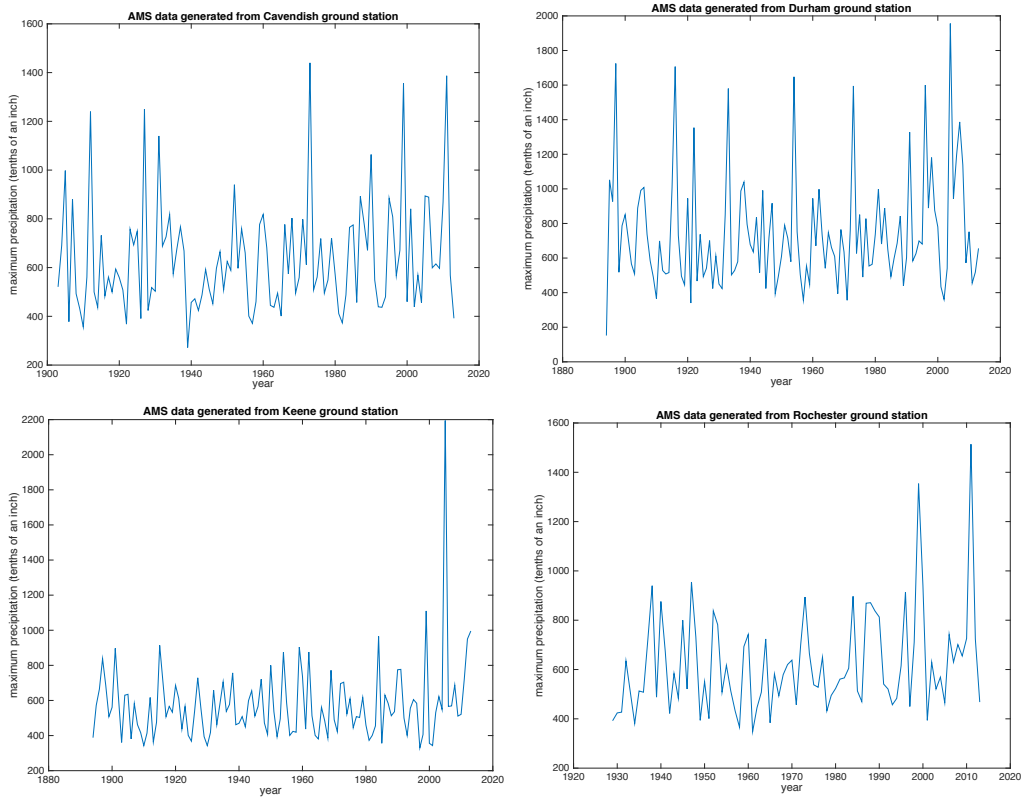
One other unanswered question is whether we can relate the IMFs to a some non-linear system. For example, could we use significant IMFs in any way to identify a set of non-linear equations that represent the driving forces behind the mode functions. While this would be difficult when we do not find significant IMFs, the other direction of this relation may be just as informative. Suppose, in our study of extreme precipitation, we were able to find a set of differential equations that we believe drive, at least in part, the precipitation on a global level (perhaps based on air movement in the atmosphere, thermodynamics, etc.). Would we be able to use this to better look for patterns in IMFs, comparing the patterns we find at different stations? How would this change the way we look at trends?

In conclusion, EMD is a powerful tool, despite the many ways it needs refinement. Unfortunately, it would seem that the noise present in the AMS data gathered from Vermont and New Hampshire stations is simply too great for us to pull out any significant IMFs, and yet the residual functions hint that there may be something at time scales too large for the scope of our data. From the scale of these residuals, one would guess that 200 or even 300 years of data should be enough to show interesting results, and we eagerly await a data set for which this is possible.

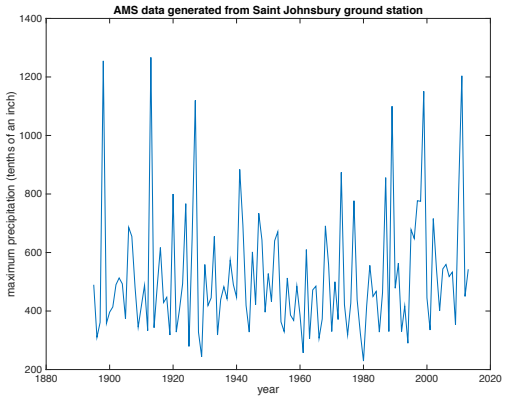
Appendix A

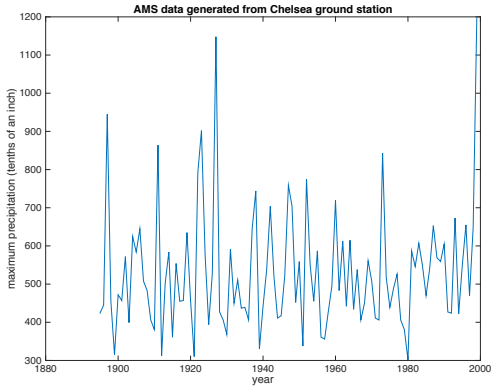
Appendix

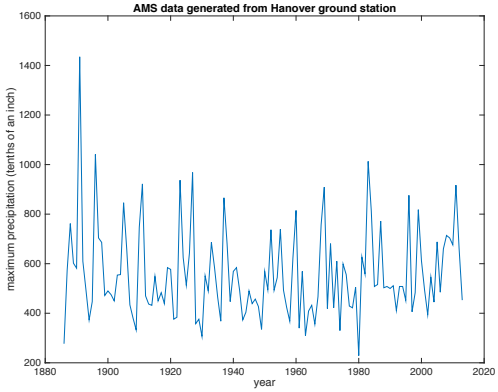
A.1 Original time series

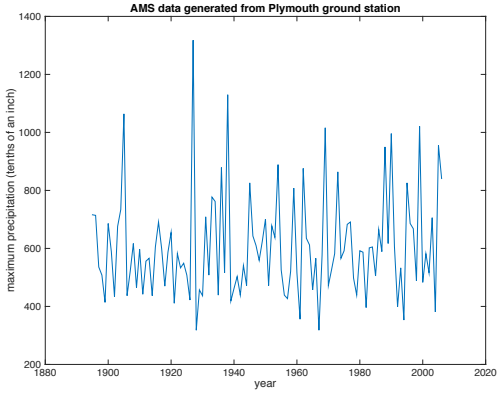


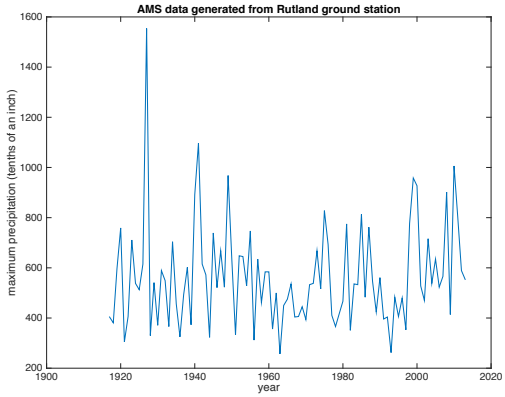
A.1. ORIGINAL TIME SERIES

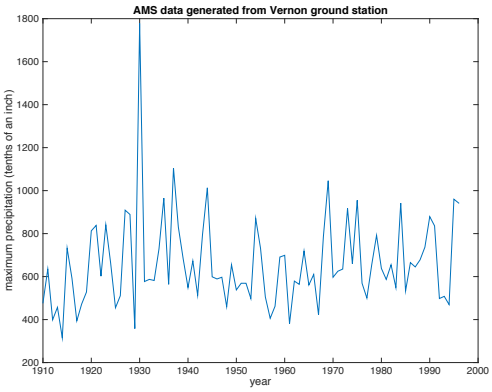






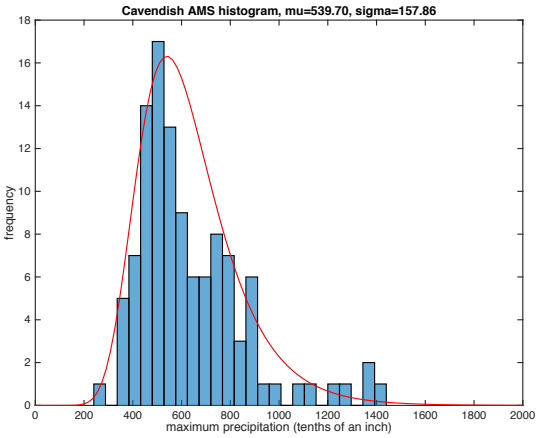


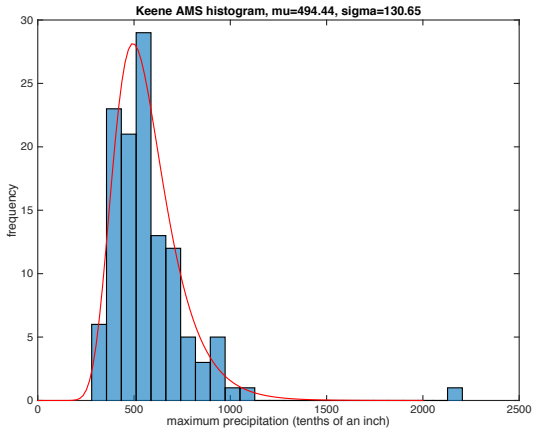


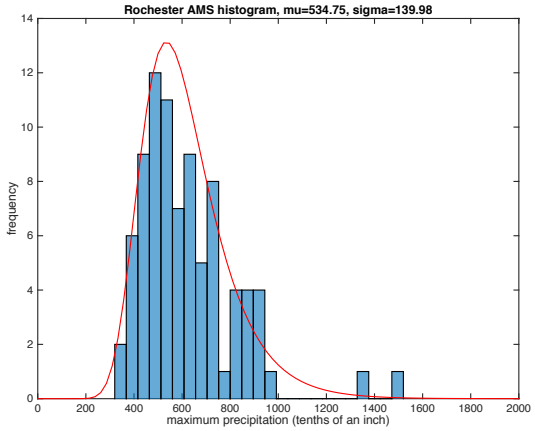


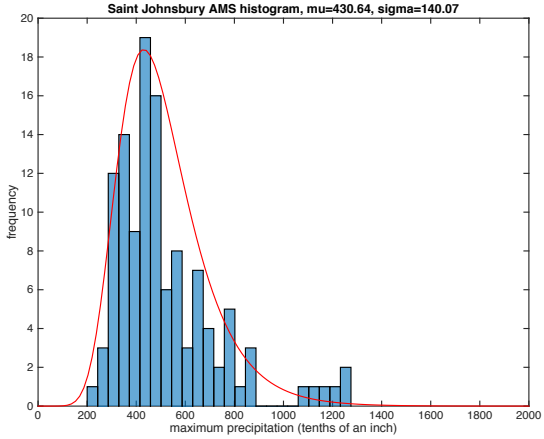
A.2 Histograms and fitted EV distributions

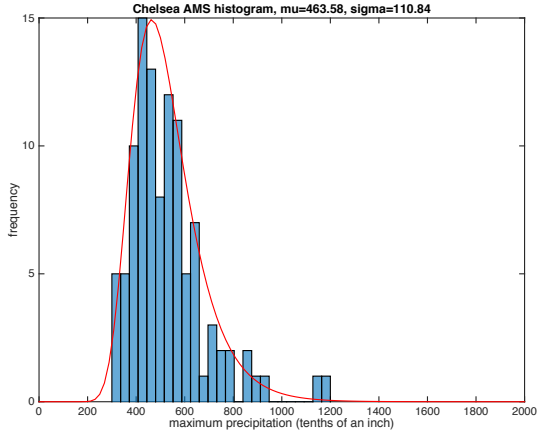
More examples of the stationary analysis of the AMS data, for each of the studied stations:

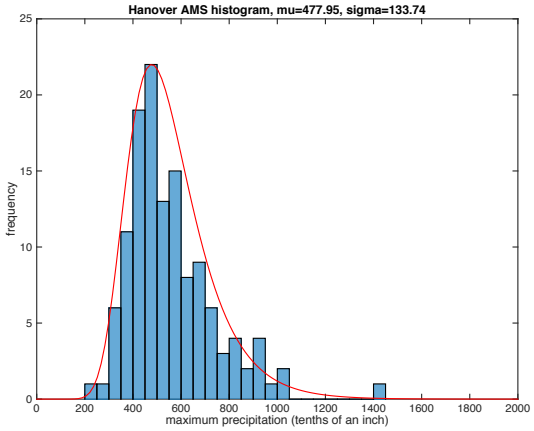


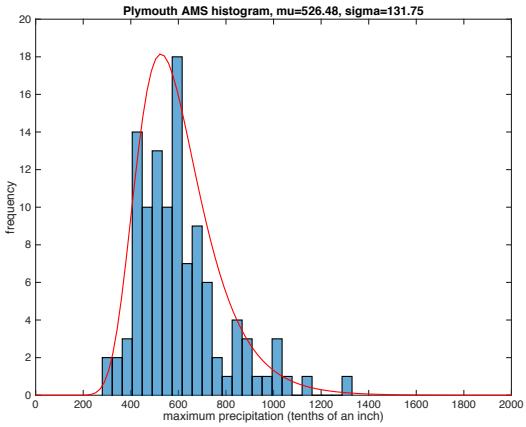


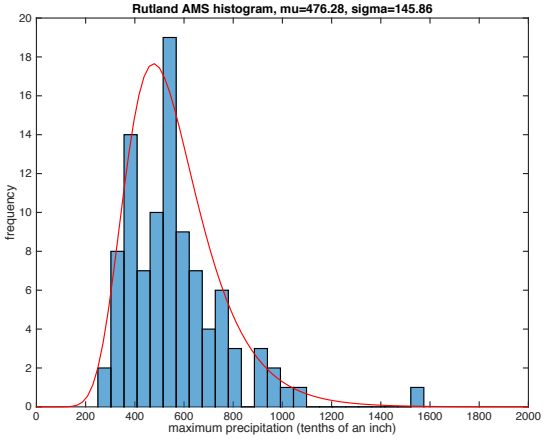


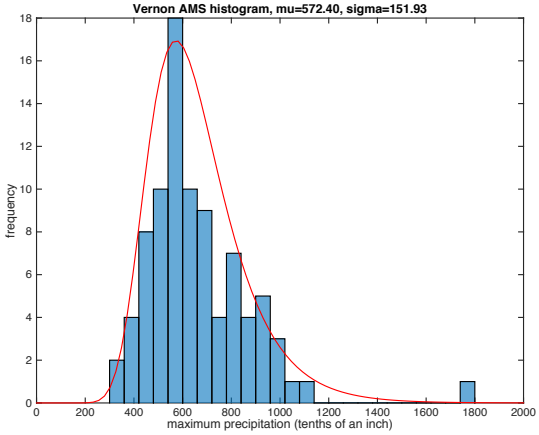





Hanover AMS histogram, mu=477.95, sigma=133.74
frequency
maximum precipitation (tenths of an inch)

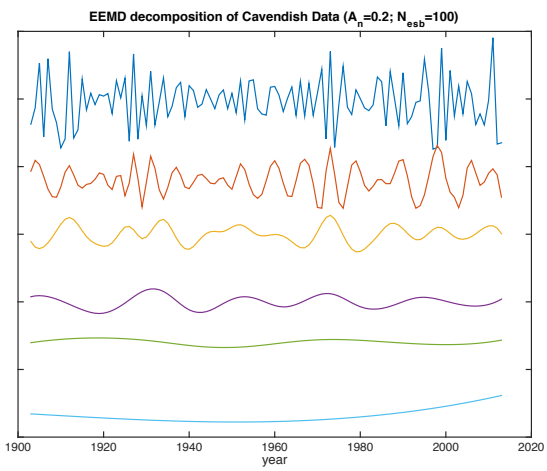

Plymouth AMS histogram, mu=526.48, sigma=131.75
frequency
maximum precipitation (tenths of an inch)

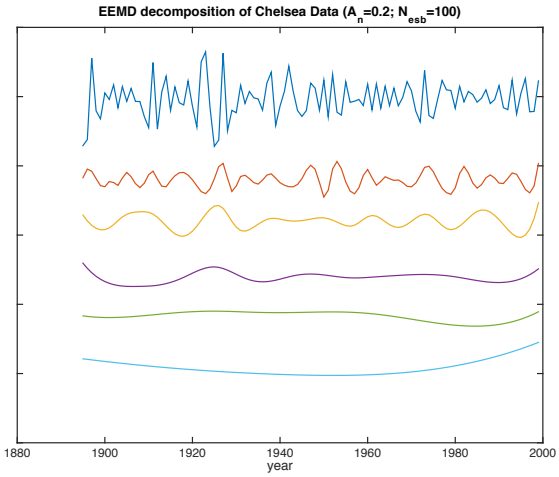

Rutland AMS histogram, mu=476.28, sigma=145.86
frequency
maximum precipitation (tenths of an inch)

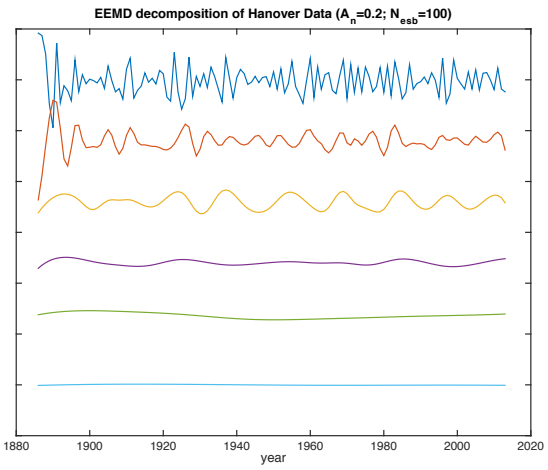

Vernon AMS histogram, mu=572.40, sigma=151.93
frequency
maximum precipitation (tenths of an inch)

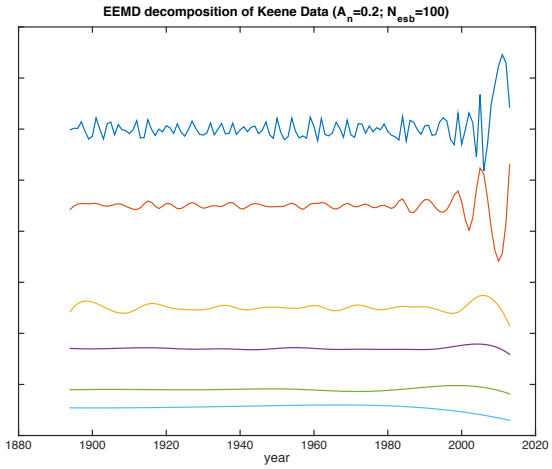
A.3 IMFs

The IMFs generated from each of the ground station time series. As noted, none of the IMFs (other than the few shown earlier) showed statistical significance:

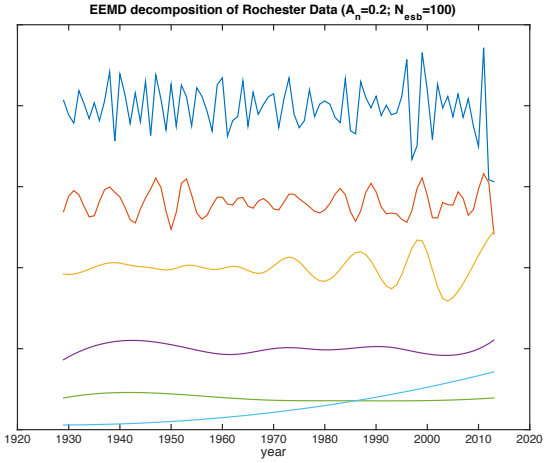


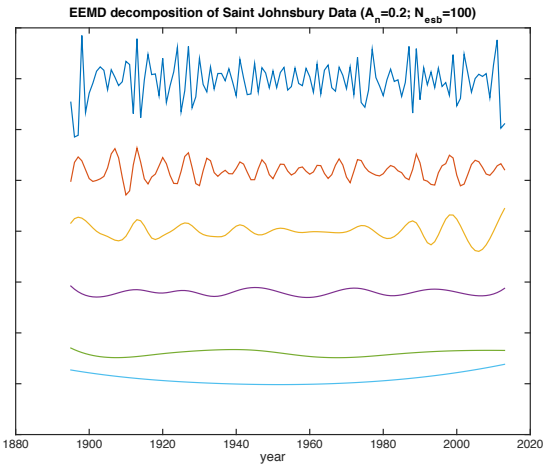


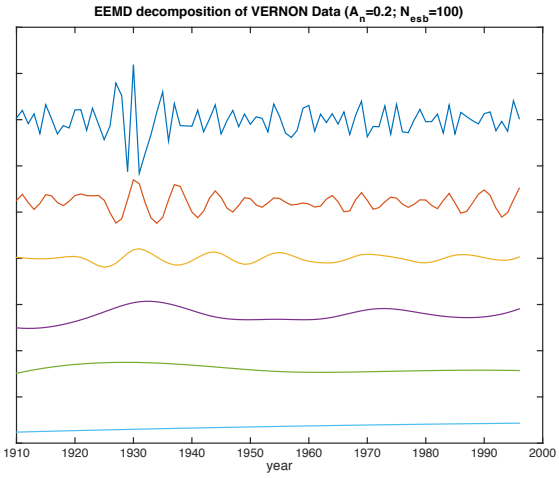




A.3. IMFS


EEMD decomposition of Rochester Data (A_n=0.2; N_esb=100)
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
year


EEMD decomposition of Saint Johnsbury Data (A_n=0.2; N_esb=100)
1880
1900
1920
1940
1960
1980
2000
2020
year


EEMD decomposition of VERNON Data (A_n=0.2; N_esb=100)
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
year

Bibliography

[1] Alex J. Cannon. A flexible nonlinear modeling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes*, 24:673–685, 2010.

[2] Alex J. Cannon and Ian G. McKendry. A graphical sensitivity analysis for statistical climate models: applications to indian monsoon rainfall prediction by artificial neural networks and multiple linear regression models. *International Journal of Climatology*, 22:1687–1708, 2002.

[3] Stuart Coles. *An introduction to Statistical Modeling of Extreme Values*. Springer: London, 2001.

[4] Matt J. Fischer and Adrian W. Paterson. Detecting trends that are nonlinear and asymmetric on diurnal and seasonal time scales. *Climate Dynamics*, 43:361–374, 2014.

[5] Norden E. Huang. http://rcada.ncu.edu.tw/research1.htm, 2014 (last accessed Febuary 16, 2015).

[6] Norden E. Huang, Zhen Shen, and Steven R. Long. A new view of non-linear water waves – the hilbert spectrum. *Annual Review of Fluid Dynamics*, 31:417–457, 1999.

[7] Norden E. Huang, Zhen Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Liu. The epirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings: Mathematical, Physical, and Engineering Sciences*, 454(1971):903–995, 1998.

[8] Norden E. Huang and Zhaohua Wu. A review on hilbert-huang tranform and its application to geophysical studies. *Reviews of Geophysics*, 46:2007RG000228, 2008.

[9] Christoph Marty and Julliette Blanchet. Long-term changes in annual maximum snow depth and snowfall in switzerland based on extreme value statistics. *Climatic Change*, 111:705–721, 2012.

[10] MATLAB. *version 8.4.0*. The Mathworks Inc., Natick, Massachusetts, 2014.

[11] NOAA. Climate data online. http://www.ncdc.noaa.gov/cdo-web/, 2014 (last accessed Febuary 16, 2015).

[12] G. G. S. Pegram, M. C. Peel, and T. A. McMaron. Empirical mode decomposition using rational splines: an application to rainfall time series. *Proceedings: Mathematical, Physical, and Engineering Sciences*, 464(2094):1483–1501, 2008.

[13] Sanja Perica, Deborah Martin, Sandra Pavlovic, Ishani Roy, Micheal St Laurent, Carl Trypaluk Dale Unruh, Micheal Yekta, and Geoffrey Bonnin. Annual maximum series and trend analysis. *NOAA Atlas 14*, 9:A.2, 2013.

[14] Yuan Shi, King-Fai Li, Yuk L. Yung, Hartmut H. Aumann, Zuoqiang Shi, and Thomas H. Hou. A decadal microwave record of tropical air temperature from amsu-a/aqua observations. *Climate Dynamics*, 41(5-6):1385–1405, 2013.

[15] Zhaohua Wu and Norden E. Huang. A study of the characteristics of white noise using the empirical mode decomposition. *Proceedings of the Royal Society of London*, 460:1597–1611, 2004.